DOCUMENT RESUME

ED 069 670 TM 002 121

AUTHOR	Andrulis, Richard S.
TITLE	Construct Validation of A Standardized Achievement Test.
PUB DATE	72
NOTE	14p.; Paper presented at meeting of the American Psychological Association (80th, Honolulu, Hawaii, Sept., 1972)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Academic Achievement; *Achievement Tests; College Students; Educational Research; Factor Analysis; Item Analysis; *Predictive Ability (Testing); Research Methodology; Speeches; Statistical Analysis; Student Evaluation; Test Construction; Tests; *Test Validity
IDENTIFIERS	Chartered Life Underwriter

ABSTRACT

The purpose of the investigation was to determine the construct validity of a standardized achievement test. The test, administered to over 5800 subjects, is one examination in a diploma program for students pursuing the Chartered Life Underwriter (CLU) designation. Results of factor and multiple discriminant analysis indicated the presence of five content and cognitive constructs. However, only 12% of the variance was accounted for by these constructs. Subsequent analysis has indicated the presence of an item response format construct that might relate with individual performance. (Author/DJ)

ED 069670

CONSTRUCT VALIDATION OF A STANDARDIZED ACHIEVEMENT TEST

Richard S. Andrulis

American College of Life Underwriters

The purpose of the investigation was to determine the construct validity of a standardized achievement test. The test, administered to over 7800 subjects, is one examination in a diploma program for students pursuing the Chartered Life Underwriter (CLU) designation.

Results of factor and multiple discriminant analysis indicated the presence of five content and cognitive constructs. However, only 12% of the variance was accounted for by these constructs. Subsequent analysis has indicated the presence of an item response format construct that might relate with individual performance. Further investigation, including computation of the shared variance between item response formats and content traits, is underway.

TM 002 121

FILMED FROM BEST AVAILABLE COPY

Test and Population Parameters

Before exploring the procedures used to investigate the research questions, the characteristics of the test and population involved in the study will be enumerated.

Twice a year, The American College of Life Underwriters administers examinations covering a total of ten courses. Five objective course examinations are given in January and repeated in June, along with objective and essay examinations covering the other five courses.

The examination chosen for study was Test I for Course Area One. The rationale for the selection of this examination was threefold: (a) it is the examination that most students take first, (b) it is administered in January and June and has a higher rate of failure than the other nine examinations and (c) personnel in other developmental projects not associated with this investigation were seeking information about its nature and value.

The first test has 100 multiple choice questions divided into five subtests of 12 to 27 items each. The items are arranged into subtests according to subjective judgment on content homogeneity, without the guidance of statistical information. Further, three response formats are used in the examination with items for each format grouped together. These response formats contain items from several or all of the five subtests. The first response format (Response Format I) provides a stem and five alternative choices lettered A to E. The second format (Response Format II) provides a stem that includes three to five statements. The response to this type of item demands the selection of one, two, or a combination of all statements as the correct response. (Thorndike and Hagen, 1969, Chap. 4) The third item format (Response Format III) provides a stem with the phrase "and all of the following are true except . . ."

This final response format demands that the student select the wrong response within the context of the stem statement.

The population used in this investigation are all students in the CLU Diploma Program who completed Test I in January of 1971. The examination was administered to 5,834 individuals in test centers throughout the United States. For a large majority of the students, this was the first CLU examination.

Procedures

The tests were scored and analyzed by the Educational Testing Service (ETS) during February of 1971. The initial statistical analysis by ETS yielded item and test statistics, i.e., item difficulty. biserial correlation of items with total test score, total test score means, variances, skewness and kurtosis measures, as well as split-half reliability coefficients.

Subsequent to the analysis by ETS, the data were further analyzed by the staff at the American College during the summer of 1971 according to the following procedures:

I Duplication of the initial test analysis to determine item and test characteristics.

II Factor Analysis of the entire examination and each subtest separately, using a Principal Components method with a Varimax Rotation.

III Multiple Discriminant Analysis of the entire test, performed by dividing the subjects' total test scores into 3 groups -- the top and bottom 27% and the middle 46%

I The item and test statistics computed for the examination were as follows: Test means, standard deviation and reliability (Cronbach's Alpha coefficient) for each of the five subtests and the total test; item difficulty levels, point-biserials coefficient of correlation of each item with the total and

subtest scores; and finally, a distribution of the alternative responses chosen by the examinees.

II Two factor analyses were performed. The first was on all 100 items collectively. An eigenvalue of 1.0 was used as the criterion for stopping factor extraction. The second phase of the analysis was a factor analysis of each individual subtest. As with the initial factor analysis, the criterion for stopping factor extraction was an eigenvalue of 1.0.

III The multiple discriminant analysis was done using all 100 items. Dividing the test scores into the three groups defined previously, statistical analysis identified itmes that significantly discriminated among the three levels of performance.

Results

Item Analysis

The results of the first phase of the analysis are reported in Table I.

Insert Table I about here

The subtest and total test score distributions tend to be negatively skewed and mildly leptokurtic. With the exception of subtest one, the reliability tended to be generally good, especially for the total test. Further, the item point-biserial coefficients of correlation with the total test scores generally averaged in the .35 - .45 range; and .45 - .50 with the subtest score. The general indication is that the test contains homogeneous items of similar content for each of the subtests as well as for the total test.

However, analysis of the item data, categorized according to the three response formats, pointed to a possible problem area. These data are presented in Table II.

Insert Table II about here

There appeared to be a difference in test performance that might be dependent upon the response format used in the examinations. However, this tentative conclusion demanded further investigation.

Factor Analysis Results

The next phase of the analysis was the extraction of factors from the total test. The factor analysis yielded 24 factors that accounted for 37.3% of the trace. Five factors were clearly identifiable as related to specific content areas. These are presented in Table III.

Insert Table III about here

Factors one, two and three were tentatively identified as having mathematic dimensions, reflecting comprehension of mathematical definitions, calculation and problem solving aspects respectively. The items loading on each of the three factors tended to have high intercorrelations, which in addition to the content similarity, led to the conclusion of a predominant mathematics factor. However, it was discouraging that only 7% of the variance was accounted for by the three factors. The fourth and fifth factors, when reflected, were interpreted as primarily verbal operations with mathematical overtones. The fourth factor extracted was described as a comprehension of insurance definitions and the fifth, as a comprehension and application of insurance principles within a mathematics domain. Surprisingly, these two factors accounted for 5% of the variance.

In the context of these results, the obvious question is what occurred to the distribution of the remaining variance. A total of 24 factors were extracted, but only 5 were clearly recognizable as having a specific content and cognitive feature. Of the remaining nineteen factors, many had loadings from one or two items in extremely specific content areas. However, there were two factors that reflected not only a content area, but also appeared to reflect a test characteristic. The test characteristic appeared to be the response formats that were used throughout the examination. These are data presented in Table IV.

Insert Table IV about here

Items loading on Factor 3 had characteristics of Response Format II, while items loading on Factor 15 had characteristics of Response Format I. Even though all items that loaded on Factors 3 and 15 respectively were of the two response formats, it was difficult to separate the contribution of the content and format to this factor pattern. For one thing, the number of items that loaded on the factors was not extensive, even though the loading was relatively clear. Secondly, there were still weak content similarities that existed within the items that loaded on Factors 3 and 15 respectively.

Further exploration was necessary to determine the extent to which the response formats might be affecting the factor structure of the test. For this reason, it was decided to continue the analysis by computing a multiple discriminant function. Dividing the total group of 5,834 individuals into three groups (top 27% of total scores, middle 46%, and bottom 27%), according to their scores on the total test, the multiple discriminant analysis then used each item to differentiate among the three levels of total test performance.

The findings were inconclusive. Of the 100 items that were used as the input variables, 20 items significantly discriminated among the three groups. These items had F ratios significant at $P < .05$. Examination of the 20 items in the context of the three response format revealed the information presented in Table V.

Insert Table V about here

Chi Square analysis performed on the data in Table V did not indicate that there was a significantly greater number of items of one response format over the other response formats.

However, what was more revealing was that only 13 of the 20 items that differentiated among the three groups of subjects loaded on one of the 24 factors. Seven of the most potent items did not load on any factor. This led to a tentative indication that much of the variance of the total test was, in fact, unaccounted for, because items of key content and cognitive areas were not included in the test. Of the seven items that did not load on any factor, another interesting characteristic was noted. This is reflected in Table VI.

Insert Table VI about here

A question can be raised as to whether or not further exploration of the Response Format and individual test performance might provide information on the construct validity of the test.

Conclusion

At this point, the results do not indicate a clear construct pattern to the test. The evidence of interaction between students' scores, content and response format must be explored further before any conclusions can be drawn.

TABLE I

Descriptive Statistics of the CLU Examination One Administered in January, 1971

Subtest Number	1	2	3	4	5	Total Test
Number of Items	12	21	19	27	21	100
Means	8.3	16.9	13.7	17.3	13.0	69.5
Standard Deviation	1.9	2.8	2.7	5.0	3.1	12.6
Alpha Coefficients	.49	.67	.60	.79	.62	.89

TABLE II

Total Test and Item Means of Examination One Categorized by Response Format

	No. of Items	Total Test Mean	Average Item Difficulty
Response Format I	42	28.5	.68
Response Format II	36	24.1	.67
Response Format III	22	16.5	.75

TABLE III

Factor Loadings for Five Factors Extracted from Examination One

ITEMS	Factors I	II	III	IV	V
7	53[a,b]				
8	40				
10	34				
18	48				
25	33				
29	36				
41	52				
46	39				
49	45				
60	36				
77	32				
89	34				
21		48			
43		43			
9			39		
13			47		
17			31		
40			41		
68			33		
4				-46	
12				-40	
19				-37	
23				-37	
27				-41	
31				-34	
32				-44	
35				-46	
54				-45	
56				-40	
11					-38
39					-37
42					-43
60					-54
62					-40
86					-35
97					-32

Note: (a) values in excess of .30 are reported
(b) decimal points have been omitted

TABLE IV

Factor Loading of Items from Response Formats I and II for Examination One

	Item Number	Factor 3	Factor 15
Response Format II	50	60[a,b]	
	63	49	
	76	30	
Response Format I	33		35
	40		41
	17		31
	13		47
	9		39

Note: (a) values in excess of .30 are reported
(b) decimal points have been omitted

TABLE V

Test Items that Significantly Descriminate Among Three Groups of Subjects Related to Response Formats in Examination One

Response Format Type	Number of Items that Significantly Discriminated Among the 3 Groups	Total No. of Items in the Test for Each Response Format
I	10	42
II	4	36
III	6	22

Table VI

Test Items Unrelated to any Factor that Significantly Discriminate Among Three Groups of Subject-Reported by Response Format for Examination One

Response Format Type	Distribution of the 7 Items Obtained from the Multiple Discriminant Analysis	Total No. of Items from Each Response Format
I	2	42
II	3	46
III	2	22

REFERENCES

Thorndike, R. and Hagen, E. Measurement and Evaluation in Psychology and Education, 3rd ed. New York, New York: John Wiley and Sons, Inc., 1969.

Nunnally, J. Psychometric Theory, New York, New York: McGraw Hill, 1967.

Veldman, D. Fortran Programing for the Behavioral Sciences, New York, New York: Holt, Rinehart and Winston, 1967.

Educational Testing Service Statistical Reports to The American College of Life Underwriters. In-house Reports, 1971.